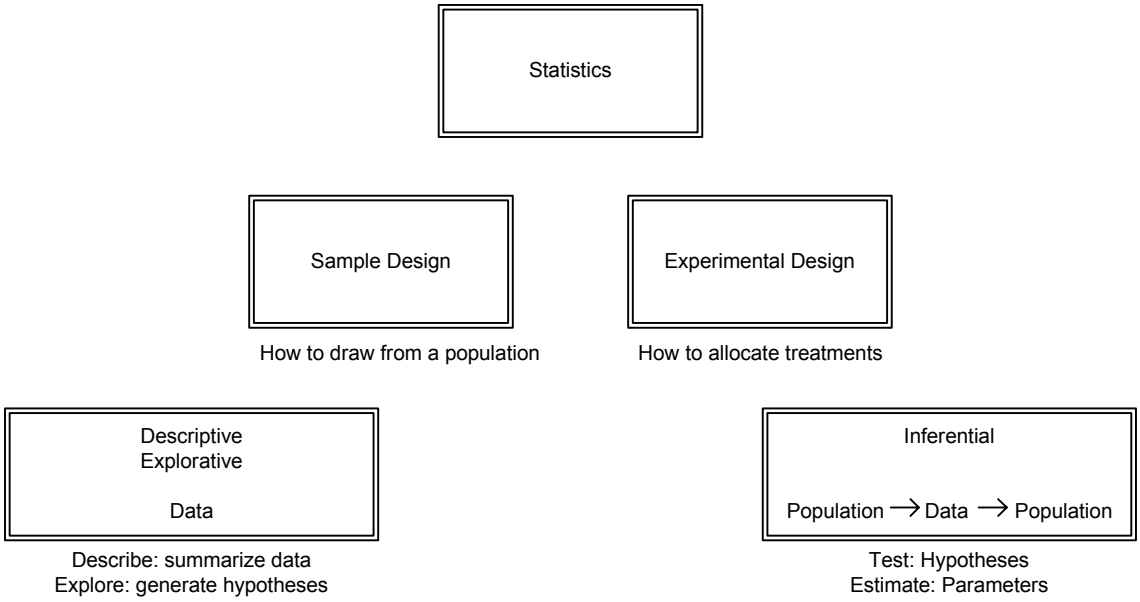
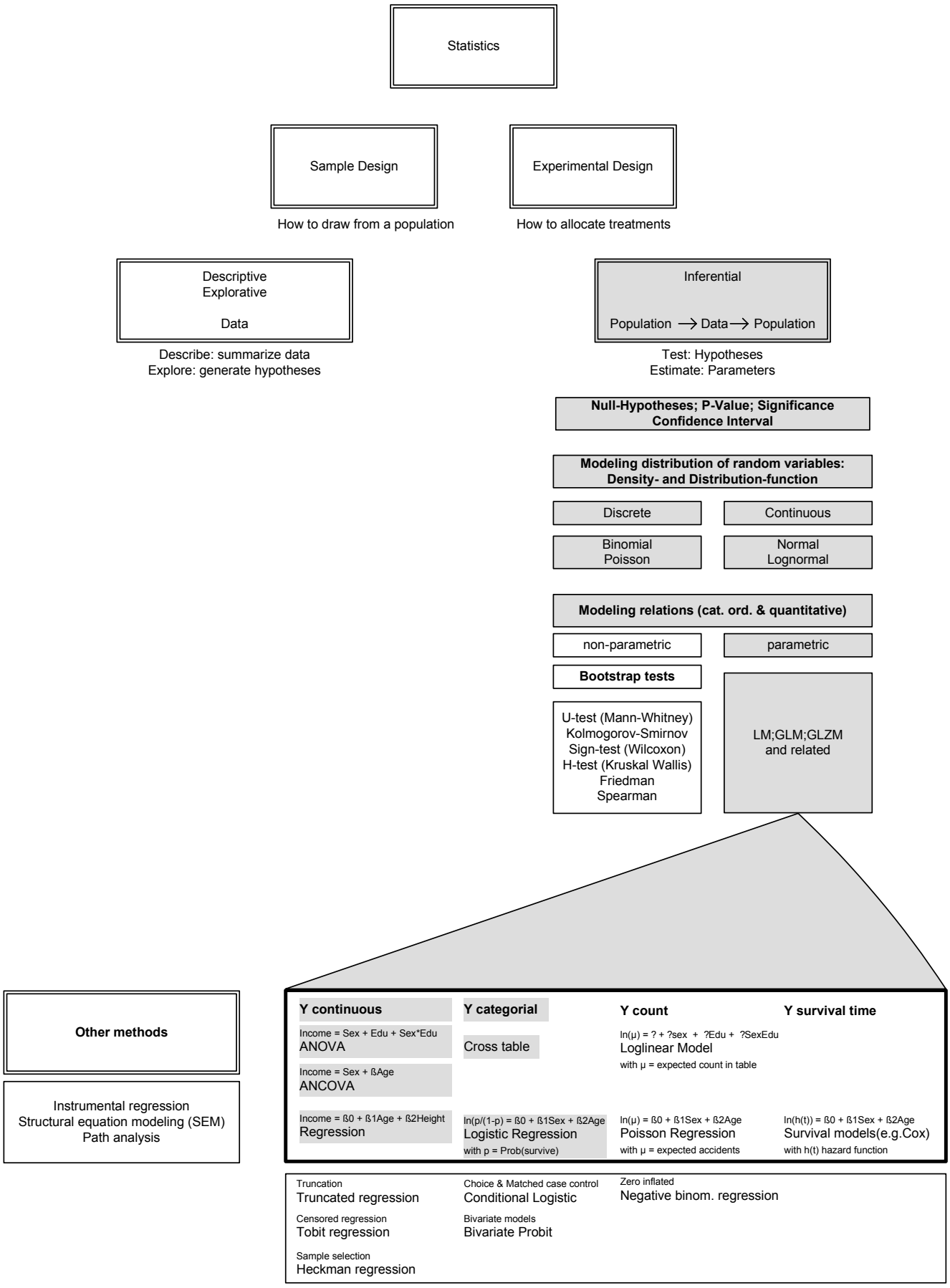
About this

Inferential Statistics

This text is intended to give an overview about definitions, theorems and rules of thumb.
It is compiled mainly from basic chapters of three textbooks: Moore (2006), van Belle (2008) and Good (2006).
The text will also inform you about the state of the arts and controversial questions.
The text is not intended to substitute the ordinary textbooks.





Grayed topics are covered in this overview

Population, Sample, Types of study[1]

All of statistical inference is based on one idea:

To see how trustworthy a procedure is, ask what would happen if we repeat it many times.

Population	The entire group of individuals that we want information about is called the population. <M>

Sample	A part of a population to represent the whole. <M>

Example of a random sample:

Simple random sample (SRS)	… of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected. <M>

Example of a non-random sample:

Voluntary response sample	… consists of people whose choose themselves by responding to a general appeal. <M>

Remark: Voluntary response samples are biased because people with strong opinions, especially negative opinions, are most likely to response. <M>

[1] Direct citations from: Moore (2006, Chapters 3.2-3.4) <M>; Utts (2004, Chapter 5) <U>; Levy PS, Lemeshow S (2008, Chapter 2) <L>

Population, Sample, Types of study[2]

A sample survey	Collects information about a population by selecting and measuring a sample from the population. <M>
Observational study	In an observational study we observe differences in the explanatory variable and then notice whether these are related to differences in the response variable. <U>
Randomized experiment	In a randomized experiment, we create differences in the explanatory variable and then examine the result. <U>

"The concept of significance level, power, p-value, and confidence interval apply only to data that have arisen from carefully designed and executed experiments and surveys." (Good, 2006, p.115)

Cause and Effect can only be inferred by randomized experiments <M>

An observational study, even one based on a statistical sample, is a poor way to gauge the effect of an intervention…. **When our goal is to understand cause and effect, experiments are the only source of fully convincing data**. <M>

The main purpose of random assignment of treatments, or the order of treatments, is to even out confounding variables across treatments.
By doing this, a cause- and-effect conclusion can be inferred that would not be possible in an observational study. <U>

[2] Direct citations from: Moore (2006, Chapters 3.2-3.4) <M>; Utts (2004, Chapter 5) <U>; Levy PS, Lemeshow S (2008, Chapter 2) <L>

Statistic (Estimator) → Parameter

Parameter	A number that describes the population. A parameter is a fixed number; but in practice we do not know its value. <M>
Statistic	A number that describes a sample. The value of a statistic is known when we have taken the sample, but it can change from sample to sample. **We often use a statistic to estimate an unknown parameter.** <M>
Sampling distribution of a statistic	The distribution of values taken by a statistic in all possible samples of the same sample size from the same population. <M>

Let $\hat{d}$ be an estimator for the parameter d

Bias of a statistic	Concerns the center of the sampling distribution. A statistic used to estimate a parameter is unbiased if the mean of its sampling distribution is equal to the true value of the parameter estimated. To reduce bias, use random sampling (like SRS). <M> $\text{Bias}(\hat{d}) := E(\hat{d}) - d$
Variability of a statistic	Describes the spread of its sampling distribution. The spread is determined by the sampling design and the sample size n. … You can make the variability as small as you want by taking a large enough sample <M> $\text{Var}(\hat{d})$
Mean square error of a statistic	How far away a particular value of the estimate is, on average, from the true value of the parameter being measured. <L p.36> $\text{MSE} := E\left([\hat{d} - d]^2\right) = \text{Var}(\hat{d}) + \text{Bias}^2(\hat{d})$ Assesses the quality of an estimator in terms of its variation and unbiasedness

Remark:
The actual error in estimating a parameter by a statistic can be much larger than the sample distribution suggests. What is worse, there is no way to say how large the added error is. (e.g. due to undercoverage or nonresponse in a sample, or a lack of realism in an experiment) <M>

Statistic (Estimator) → Parameter

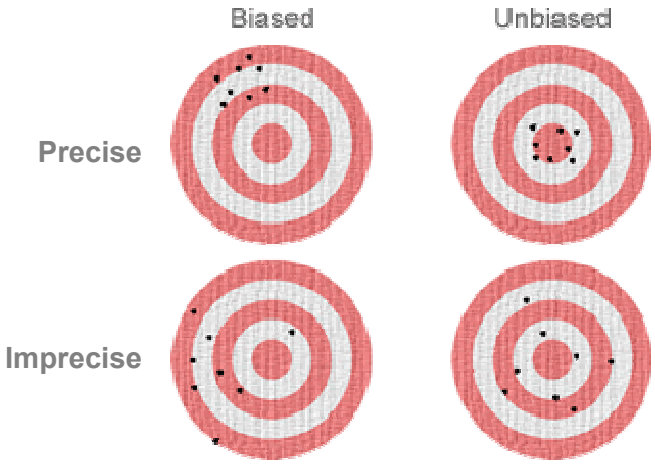
Related definitions:

Precision[3] of an estimator: $1/\mathrm{Var}(\hat{d})$

Reliability of an estimator: ~ Precision
How reproducible the estimator is over repetitions of the process yielding the estimate (can be expressed in terms of its sampling variance, or equivalently, its standard error. The smaller the standard error of an estimator, the greater is its reliability.
<L p.35>

Validity[4] of an estimator: ~ unbiased
How the mean of the estimator over repetitions of the process yielding the estimate, differs from the true value of the parameter being estimated (can be evaluated by examining the bias of the estimator: The smaller the bias, the greater is the validity <L p.35>

Accuracy[5] of an estimator: ~ MSE
How far away a particular value of the estimate is, on average, from the true value of the parameter being measured. Generally evaluated on the basis of its mean square error (MSE) or root MSE (RMSE).
$\mathrm{MSE}(\hat{d}) := E[(\hat{d}-d)^2] = \mathrm{Var}(\hat{d})+\mathrm{Bias}^2(\hat{d})$
<L p.36>



"Bias and precision of estimates
We measure the precision of an estimator by how far we expect it to be from the average of the estimates we might obtain from many similar surveys. Here it is represented by the average distance from the individual points to the center point of the cluster. This average distance is the standard error of our estimate.
Precise estimates (top row) have small standard errors and imprecise ones (bottom row) have large standard errors." PEAS (2006)

[3] "Precision" is sometimes used as confidence interval
[4] "Validity" outside this narrow statistical context addresses a wide range of questions (e.g. validity in social science in Wikipedia: http://en.wikipedia.org/wiki/Validity_(statistics))
[5] "Accuracy" is sometimes used for validity. It is also used in describing classification quality. Accuracy=100% means that all diseased and non-diseased people are classified correctly.

$\bar{x}$ is unbiased estimator for μ and SEM = Std/√n

Let $x_1,\ldots,x_n$ be a random sample sample of size n.

That is $x_1,\ldots,x_n$ are independent and identically distributed (i.i.d.) with $E(x_i)=\mu$ and $Var(x_i) = \sigma^2$

Then:

1) **The sample mean $\bar{x}$ is an unbiased estimator $\hat{d}$ for the parameter μ**
2) $Var(\bar{x}) = MSE(\bar{x}) = \frac{\sigma^2}{n}$
3) **SEM= Std/√n** if we estimate σ by the standard deviation of the sample (Std)

Proof:

1) $E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \overset{\text{generally}}{=} \frac{1}{n}E\left(\sum_{i=1}^{n} x_i\right) \overset{\text{generally}}{=} \frac{1}{n}\sum_{i=1}^{n} E(x_i) \overset{\text{identical}}{=} \frac{1}{n}\cdot n \cdot E(x) = \mu$

   so that $Bias(\bar{x}) = E(\bar{x}) - \mu = 0$ i.e. $\hat{d}$ is an unbiased estimator of μ.

2) $Var(\bar{x}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \overset{\text{generally}}{=} \frac{1}{n^2}Var\left(\sum_{i=1}^{n} x_i\right) \overset{\text{independent}}{=} \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(x_i)\right) \overset{\text{identical}}{=} \frac{1}{n^2}\cdot n \cdot Var(x) = \frac{\sigma^2}{n}$

   and $MSE(\bar{x}) = Var(\bar{x}) + Bias^2(\bar{x}) = \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n}$

   So that $RMSE := \sqrt{MSE} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{Var(\hat{d})} = \frac{\sigma}{\sqrt{n}}$

3) When we estimate σ by $Std := \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$, the standard deviation of the sample, we have the well-known formula for the standard error of the mean (SEM)

$$SEM = \frac{Std}{\sqrt{n}}$$

Note: Normally distributed is not necessary

Var is an unbiased estimator for σ^2, but Var_n is biased

Let $x_1,\ldots,x_n$ be a random sample sample of size n.

That is $x_1,\ldots,x_n$ are independent and identically distributed (i.i.d.) with $E(x_i)=\mu$ and $Var(x_i) = \sigma^2$

Then:

1) Var = $\frac{1}{n-1}\sum(x_i-\bar{x})^2$ **is a unbiased estimator for σ^2**

2) Var_n = $\frac{1}{n}\sum(x_i-\bar{x})^2$ **is a biased estimator for σ^2**

Proof:

1)
$$
\begin{aligned}
E(Var) &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] \\
&= \frac{n}{n-1}\cdot\frac{1}{n}\left\{E\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\right\} \\
&= \frac{n}{n-1}\cdot\frac{1}{n}\left\{\sum_{i=1}^{n}E\left[(x_i-\bar{x})^2\right]\right\} \\
&= \frac{n}{n-1}\cdot\frac{1}{n}\left\{\sum_{i=1}^{n}E\left[\{(x_i-\mu)-(\bar{x}-\mu)\}^2\right]\right\} \\
&= \frac{n}{n-1}\cdot\frac{1}{n}\left\{\sum_{i=1}^{n}E\left[(x_i-\mu)^2-2(x_i-\mu)(\bar{x}-\mu)+(\bar{x}-\mu)^2\right]\right\} \\
&= \frac{n}{n-1}\cdot\frac{1}{n}\left\{\sum_{i=1}^{n}E(x_i-\mu)^2-2\cdot E[(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\mu)]+\sum_{i=1}^{n}E(\bar{x}-\mu)^2\right\} \\
&= \frac{n}{n-1}\cdot\frac{1}{n}\left\{n\cdot\sigma^2-2\cdot E\left[(\bar{x}-\mu)\cdot(n\cdot\bar{x}-n\mu)\right]+n\cdot\frac{\sigma^2}{n}\right\} \\
&= \frac{n}{n-1}\cdot\frac{1}{n}\left\{n\cdot\sigma^2-2\cdot n\cdot E(\bar{x}-\mu)^2+n\cdot\frac{\sigma^2}{n}\right\} \\
&= \frac{n}{n-1}\cdot\frac{1}{n}\left\{n\cdot\sigma^2-2\cdot n\cdot\frac{\sigma^2}{n}+n\cdot\frac{\sigma^2}{n}\right\} \\
&= \frac{n}{n-1}\cdot\left\{\sigma^2-\frac{\sigma^2}{n}\right\} \\
&= \frac{n}{n-1}\cdot\left\{\frac{n\cdot\sigma^2-\sigma^2}{n}\right\} \\
&= \frac{n}{n-1}\cdot\left\{\frac{\sigma^2\cdot(n-1)}{n}\right\} \\
&= \sigma^2
\end{aligned}
$$

$$E(\bar{x}-\mu)^2 = MSE = Var(\bar{x}) + Bias^2(\bar{x}) = \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n}$$

2) because of 1)

$$E(Var_n) = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = E\left[\frac{n-1}{n}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \frac{n-1}{n}\cdot E(Var) = \frac{n-1}{n}\sigma^2$$

Note: Normally distributed is not necessary

Sample Population & Inferential Statistics

Population size doesn't matter	The variability of a statistic from a random sample does not depend on the size of the population, as long as the population is at least 100 times larger than the sample. <M>

Example:
"An SRS of size 2500 from the 220 million adult residents of the United States gives results as precise as an SRS of size 2500 from the 665,000 inhabitants of San Francisco." <M. p.238>

Inferential statistics	Drawing inferences about a population on the basis of observations obtained from a sample.<M>

P value; Test-Statistics; Significance

Categorical data

- Hypothesis: There is something wrong with the "6" of this die
- Null-Hypothesis (H0): Nothing is wrong. i.e. "6" has probability 1/6
- Testing H0: cast a die 6 times[6]

Sample-data:

6	6	6	3	3	6

4 times a "6" in 6 casts

If the Null-hypothesis is true, what is the probability of our observed data?

Probability(first 4 times a "6" and the next 2 times "no 6") = $\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^2$ = 0.00054

there are 15 combinations which would give such a result

```
 1) x x x x - -
 2) x x x - x -
 3) x x - x x -
 4) x - x x x -
 5) - x x x x -
 6) x x x - - x
 7) x x - x - x
 8) x - x x - x
 9) - x - x x x
10) x x - - x x
11) x - x - x x
12) - x x - x x
13) x - - x x x
14) - x - x x x
15) - - x x x x
```

So that: Probability(4 times a "6" in 6 casts) = 15 * 0.00054 = 0.0080

First provisional definition

P value (p) : = Probability(actual sample | H_0)

in our example:
p = Prob(4 times a "6" in 6 casts| H0) = 0.00804

If we feel or consider that the actual sample is unlikely we would say that our data are incompatible with the null-hypothesis.

[6] Draw from all possible samples of size 6

As more extreme results would also show incompatibility with H0 we should take them into account

So, to be on the safe side, we improve the definition:

Second provisional definition

P value (p) : = Probability(actual or more extreme sample | H_0)

p= Prob(4 x "6" in 6 casts| H0) + Prob(5 x "6" in 6 casts| H0) + Prob(6 x"6" in 6 casts| H0) =

0.00804 + 0.00064 + 0.00002 = 0.00870

In order to find a final definition for P value we should consider that counting "6" is just one possible aspect of our sample.

Counting the "6" is a **test statistic T** of our actual sample.

For every H0 we can choose an appropriate test-statistics.

Example S* | 6 | 6 | 6 | 3 | 3 | 6 | =

H0	Test-Statistic T	T(S*)
P("1")=1/6	T = counting "1"	0
P("3")=1/6	T = counting "3"	2
P("6")=1/6	T = counting "6"	4
P("even")=1/2	T = counting even eyes	4
…	…	…

Final Definition of P value

P value (p) : = Probability(T(S) ≥ T(S*) | H_0)

The probability that the test statistic from a sample would take a value

as extreme as or more extreme than that actually observed under the assumption that H0 is true,

The smaller the P value,

the stronger is the evidence against H0

the more incompatible are data and H0.

Definition of significance level α

We can compare the P value with a fixed value that we regard as decisive.
This decisive value is called the **significance level α**.
Historical values are α=0.05; 0.01; 0.001

Significance in statistic is used to indicate only that the evidence against the null hypothesis reached the standard set by α.

Reporting the p-value[7]

Forms that should be avoided:

"The results were statistically significant."; "The results were statistically significant (*)";
"The mean difference was 1200 $ (**)" ; " The mean difference was 1200 $ (p<0.01)"

As the P value is more informative than a statement of significance because we can then assess significance at any level we choose, we should avoid statements like (p<0.01) and so leaving the decision to the reader whether this is enough evidence against the null-hypothesis.

"A P value is more informative than the reject-or-not finding at a fixed significance level." <M>

Suggested forms:

"We significantly (p = 0.0087) reject the hypothesis that the "6" has probability 1/6."
"The data are statistically significant at significance level 0.01 (p=0.0087)"
"The results were statistically significant (p=0.0087)"
"Data are not compatible with the null hypothesis (p=0.0087)

"A P value is a continuous measure of the compatibility between a hypothesis and data." <R>
"The P value is equivalent to the distance between the hypothesis and the statistic." <B>

Important:

"One of the most common naive misinterpretations of P values is that they represent probabilities of test hypotheses" <R>

[7] Citations: Moore (2006, p.415) <M>; Rothman (2008, pp.152-153) <R>; van Belle (2007, p. 18) <B>

P value; Test-Statistics; Significance

Quantitative data (one sample t-test)

Example[8]**:**

Hypothesis: The concentration of an active ingredient is not 0.86% (how it should be)

Assumption Concentration is normal distribution with unknown standard deviation.

Null hypothesis: The concentration of an active ingredient is 0.86% H0: $\mu = 0.86\%$

Alternative hypothesis: The concentration of an active ingredient is not 0.86% HA: $\mu \neq 0.86\%$

Testing H0: Take 3 specimens randomly from the production

Test-statistic: A test-statistic T which is often used (one sample t-test): $T = \frac{|\bar{x} - \mu|}{SEM}$

Where μ is the value of the H0 and SEM is the standard error of the mean: $SEM = \frac{Std}{\sqrt{n}}$

The probability is calculated using the t-distribution with n-1 degrees of freedom.

Result:

Sample (S*) 0.8403; 0.8363; 0.8447

$\bar{x}$ = 0.8404; Std = 0.0042; SEM = 0.0042/√3 = 0.0024;

$T = \frac{|\bar{x} - \mu|}{SEM} = \frac{|0.8404 - 0.86|}{0.0024} = 8.2$

The probability to have a sample whose test-statistic T is as extreme as or more extreme than the actual sample:

$Prob(T(S) \geq T(S^*)) = Prob(T(S) \geq 8.2) = 2(1 - \text{t-distribution}_2(8.2)) = 0.015$

Conclusion:

We reject the null hypothesis at the 5% level (p=0.015) that the concentration in the population is 0.86%.

Data are not compatible (p=0.015) with the null hypothesis that the concentration in the population is 0.86%

[8] slightly modified Moore (2006, p.413)

P value; Test-Statistics; Significance

Quantitative data (two sample t-test)

Example:

Hypothesis: The income tor male and female differ

Assumption Income for male and female are normal distributed with unknown and unequal standard deviation.

Null hypothesis: The income tor male and female are equal H0: $\mu_{male} = \mu_{female}$

Alternative hypothesis: The income tor male and female differ HA: $\mu_{male} \neq \mu_{female}$

Testing H0: Draw n_1=10 male and n_2=10 female specimens randomly from the production

Test-statistic: A test-statistic T which is often used (two sample t-test): $T = \frac{|\bar{x}_1 - \bar{x}_2|}{SEMD}$

Where SEMD is the standard error of the mean difference: $SEMD = \sqrt{S_1^2/n_1 + S_2^2/n_1}$

The probability is calculated using the t-distribution with

$\nu = \frac{[s_1^2/n_1 + s_2^2/n_2]^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$ degrees of freedom.

Result:

Sample (S*) male: 1800, 1600, 1900, 1900, 1400, 2200, 2200, 2000, 2100, 1900
female: 2300, 2000, 2200, 2200, 2200, 2300, 1800, 1900, 2100, 1800

$\bar{x}_{male}$ = 1900; Std_{male} = 254; $\bar{x}_{female}$ = 2080; Std_{male} = 193;

$$T = \frac{|\bar{x} - \mu|}{SEMD} = \frac{|1900 - 2080|}{100.9} = \frac{180}{100.9} = 1.8$$

$$\nu = \frac{[s_1^2/n_1 + s_2^2/n_2]^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)} = \frac{103561152.3}{1541653.8} = 16.8 \sim 17$$

The probability to have a sample whose test-statistic T is as extreme as or more extreme than the actual sample:

$Prob(T(S) \geq T(S^*)) = Prob(T(S) \geq 1.8) = 2(1 - \text{t-distribution}_{17}(1.8)) = 0.09$

Conclusion:

We cannot reject at the 5% level (p=0.09) the null-hypothesis that the mean income for male and female are equal.

Data are compatible (p=0.09) with the null hypothesis that the mean income for male and female are equal.

P value; Test-Statistics; Significance

Properties of tests

For large n everything becomes significant[9]

Example 1

Assume you have a sample with Std = 10 and $\bar{x}$ **=110**

H0: μ=100

$$T = \frac{|\bar{x} - \mu|}{SEM} = \frac{|110 - 100|}{\frac{10}{\sqrt{n}}} = \frac{10}{\frac{10}{\sqrt{n}}} = \sqrt{n}$$

Sample size n	T	P value
3	1.7	0.22
7	2.6	0.04
12	3.5	0.01

Example 2

Assume you have a sample with Std = 10 and $\bar{x}$ **=100.1**

H0: μ=100

$$T = \frac{|\bar{x} - \mu|}{SEM} = \frac{|110 - 101|}{\frac{10}{\sqrt{n}}} = \frac{1}{\frac{10}{\sqrt{n}}} = 0.1 \cdot \sqrt{n}$$

Sample size n	T	P value
3	0.17	0.89
7	0.26	0.80
12	0.35	0.74
270	1.64	0.10
384	1.96	0.05
663	2.58	0.01

[9] correct: If n increases even smallest differences become statistically significant

P value; Test-Statistics; Significance

Properties of tests

Did we use the appropriate test?

Example 3

Suppose that for Example 1 n=7 our data are:

92; 104; 106; 113; 116; 119; 120

Then Std = 10 and $\bar{x}$=110 and T = 2.6 which leads to p = 0.04

Now exchange value 120 with an outlier 1000

92; 104; 106; 113; 116; 119; **1000**

Then Std = 337 and $\bar{x}$=236 and T = 1.1 which leads to p = 0.32

While the original data are normal distributed the exchanged data have extreme irregularities

So that the t-test is no longer appropriate.

Remark:

By adding an extreme value the mean increases.
Therefore we should expect that the test becomes "more significant".
But because the standard deviation increases to a much larger extent the t-test tends to be more "non significant".
If only very few extreme outliers are present that is a conservative error of the t-test.

Substitute more outliers and observe the P value (via EXCEL)

P value; Test-Statistics; Significance

Significant Non-significant

Non-significance

If a test is non-significant it could be that

1. the sample size was too small
2. you used the wrong test (assumptions are violated)
3. H0 is really true

From a non-significant test you can not conclude that the hypothesis is true

Don't fall into the trap by arguing:

"This test was non-significant. But this was **only** because n was too small. Otherwise it would have been significant".

The consequence of such an argument would be that all scientist use sample size n =3.

The correct conclusions from a non significant test are:

I cannot reject at the 5% level (p=0.09) the null-hypothesis that the mean income for male and female are equal.

Data are compatible (p=0.09) with the hypothesis that the mean income for male and female are equal.

P value; Test-Statistics; Significance

Significant Non-significant

Non-significance cont.

Absence of evidence is not evidence of absence

If you decide to use hypothesis testing rather than estimating there is another trap you might fall into.

Example

Consider the following two situations (H0: mean income male = mean income female)

	Mean difference	95% confidence interval[10] for the mean difference	P value
Study 1	-1 $	[-3; 1]	p= 0.12
Study 2	-100 $	[-250; 50]	p= 0.12

Both tests are not significant (p=0.12) at significance level 0.05.

But it is quite clear that your conclusion and advice for further research should differ.

Study 1

Even with a larger sample size the absolute difference between male and female will still be less than 3$ with 95% confidence.

Conclusion: Not significant (p=0.12) and no evidence of relevant effects as the mean difference is, with 95% confidence, absolute smaller than 3$

Study 2

Although not significant, the confidence interval indicates that there might be a substantial difference.

Conclusion: Not significant but compatible with substantial effects, because the mean income for men could be down to 250$ less or up to 50$ more (with 95% confidence)

Please read:

Moore (2006) Chapter "Use and abuse of Tests" p. 424-427 especially "Don't ignore lack of significance" Rothman (2008) Chapter 10 p.151-163

The header of the chapter refers to a discussion in the prestigious New England Journal of Medicine where a researcher oversaw that, although his test was not significant, the confidence interval indicates that either a substantial risk or a substantial benefit could come from a specific HIV prophylaxe. See Moore (2006, p. 425) for further references.

[10] The confidence interval covers the real difference with 95% (see next chapter).

P value; Test-Statistics; Significance

Significant Non-significant

Significance

If a test is significant you have to decide whether the result is relevant

In Example 2 (Properties of tests), we saw that even small differences like 100 and 100.1 will become significant if n is large enough.

But, whether or not 0.1 is a relevant difference is not a question of statistics.

Also for significant test the general advice is that you should add confidence intervals (if available) to estimate the relevance of you findings.

P value; Test-Statistics; Significance

Two sided or one sided tests?

Two sided or one sided tests (or P values)

In the above examples we chose:

H0: $\mu = 0.86\%$
HA: $\mu \neq 0.86\%$

H0: $\mu_{male} = \mu_{female}$
HA: $\mu_{male} \neq \mu_{female}$

The resulting tests and P values are called **two-sided,** and they imply the conservative approach that one lets nature disprove us in two directions: either income of male are higher or lower than expected.

Sometimes (especially in older articles) you will find one sided hypotheses like

H0: $\mu \leq 0.86\%$
HA: $\mu > 0.86\%$
or
H0: $\mu_{male} = \mu_{female}$
HA: $\mu_{male} > \mu_{female}$

The resulting tests and P values are called **one-sided,** and they have **P values which are generally smaller than P values of two sided tests (half of two-sided P values).**

You should hesitate to use one-sided tests

1. With very view exceptions - e.g. when no two-sided test exists - there is practically no argument to test one-sided. So that the reader might suspect that the one-sided tests was used just because you needed significant results

2. There is also a practical problem. Because "It is cheating to first look at the data and frame HA to fit what data show" (Moore, 2006: p. 403) you might have problems to convince the reader that you did not look at the data

In the Statistical Encyclopedia (Armitage, 2005) it is summarized under

> "**One-sided vs. Two-sided P Values**
> Much has been written about how one could choose whether to cite a one- or two-sided P value. This is somewhat academic, since the *de facto* standard in the biomedical literature is for two-sided tests.
>
> … In general, the custom that most or all P values be reported as two-sided seems a good one, with the condition that if one-sided P values are used, this be indicated clearly enough so their value could be doubled by a reader.
>
> If the P values are in a range in which doubling makes a substantive difference, the evidence is equivocal enough so there will be controversy regardless of how the P value is reported."

"The use of one-sided p-values is discouraged. Ordinarily, use two sided p-values "
van Belle (2008,p.16)

Confidence interval

This chapter compiles: Utts (2004, Chapters 20-22) <U>; Moore (2006, Chapter 6, Chapter 14) <M>;

One of the most common types of inferences is to construct what is called a confidence interval.

Confidence interval	an interval of values computed from sample data that covers the true population parameter with a chosen probability

The most common level of confidence used is 95%.
<U>

The purpose of a confidence interval is to estimate an unknown parameter with an indication of how accurate the estimation is and how confident we are that the result is correct. <M, p.395>

Any confidence interval has two parts: an interval computed from the data and the confidence level.

The interval often has the form

estimate ± margin of error

Example (mean of a normal population unknown mean μ and standard deviation)
The **margin of error for the mean** μ of a normal population with unknown standard deviation, based on SRS of size n, is given by

margin of error = t * SEM

where t is the critical value of the t-distribution with n-1 degrees of freedom
which - for a confidence level 95% - could be approximated by 2.4 for n≥10 and 2 for n≥30.

So that the 95% confidence interval for the mean μ could be approximated by

$$\overline{x} \pm 2 \cdot SEM$$

Example (unknown proportion π, large-sample[11])
The **margin of error for the proportion** π, based on SRS of size n, is given by

margin of error = 2 * SE($\hat{p}$)

So that the 95% confidence interval for π could be approximated by

$$\hat{p} \pm 2 \cdot SE(\hat{p})$$

where $SE(\hat{p}) = \frac{Std}{\sqrt{n}} = \frac{\sqrt{\hat{p} \cdot (1-\hat{p})}}{\sqrt{n}}$

[11] number of success and the number of failures are both at least 15

Confidence interval

Examples

Example:

Income is assumed to be normally distributed with unknown Mean μ and Standard deviation σ

Sample:

n = 49
$\bar{x}$ = 1000 $
Std. = 140 $
SEM = Std/√n = 140 $ / 7 = 20 $

95% confidence interval: $\bar{x} \pm 2 \cdot \text{SEM}$ = 1000 $ ± 2*20

We conclude from our sample:

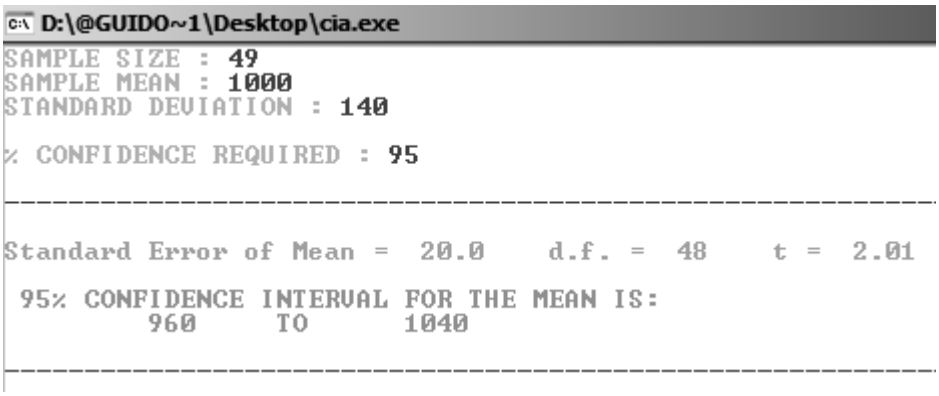
- The mean income μ of the population is estimated as 1000 $
- The 95% confidence interval for the mean income μ is [960 $; 1040 $]

Remark:

Contrary to our assumption above, we should expect that distributions of income are skewed because small values appear more often than large values.
On the other hand, the interval rely only on the distribution of $\bar{x}$, which even for quite small sample sizes is much closer to normal than that of x itself.
When n≥15, the confidence level is not greatly disturbed by nonnormal populations unless extreme outliers or quite strong skewness are present. <M, p.393>

You could also use programs like confidence interval analysis (CIA) by Altman (2004) or STATA to calculate confidence intervals immediately

CIA



STATA

```
cii 49 1000 140

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |         49        1000          20        959.7873    1040.213
```

Confidence interval

Examples

Example: <U>

From 120 volunteers randomly assigned to use a nicotine patch, 55 of them had quit smoking after 8 weeks. We use this information to estimate the probability π that a smoker recruited and treated in an identical fashion would quit smoking after 8 weeks:

$\hat{p} = 55/120 = 0.46$

$\text{Std} = \sqrt{\hat{p}(1-\hat{p})} = \sqrt{0.46 \cdot 0.54} = 0.498$

$\text{SE}\hat{p} = \text{Std}/\sqrt{n} = 0.498/\sqrt{120} = 0.498/10.95 = 0.045$

95% confidence interval: $\hat{p} \pm 2 \cdot \text{SE}\hat{p}$ = 0.46 ± 2 · 0.045 = [0.37 ; 0.55]

We conclude from our sample:

- The probability that a smoker would quit smoking after 8 weeks is estimated as 0.46
- The 95% confidence for this probability is [0.37 ; 0.55]

Remark:

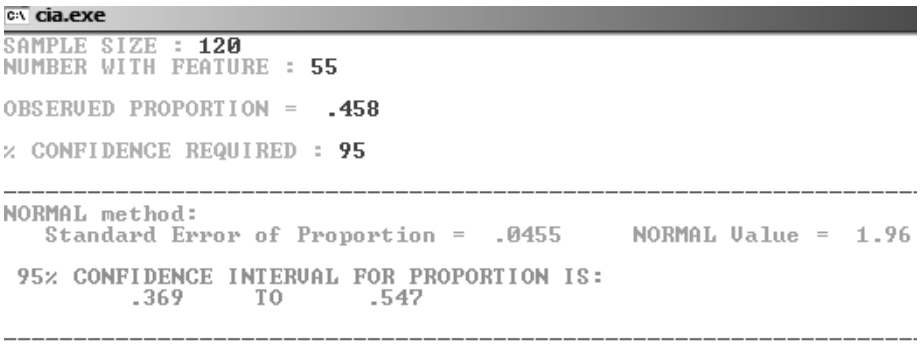
A sample size sufficiently large to use the approximation of the last example is defined as: number of success and the number of failures are both at least 15.

If your sample size is not sufficiently large you could either

- use the "Plus four confidence interval" method (Moore, 2006, p. 539) or
- calculate exact confidence via programs like
  confidence interval analysis (CIA) by Altman (2004) or STATA

Using CIA or programs like STATA should be preferred.

CIA



STATA

cii 120 55

Variable	Obs	Mean	Std. Err.	-- Binomial Exact -- [95% Conf.	Interval]
	120	.4583333	.0454848	.3670599	.5517119

Confidence interval

Other methods to construct confidence intervals

Non-parametric (Bootstrap)

For sample sizes ≥ 10 we can calculate the confidence interval for practical all parameters via Bootstrap.

This is a new method - but only practical with the aid of computers.

The basic idea is to act as if our sample were the population.
We take many samples of it. Each of these is called a resample.
We get different results from different resamples because we sample *with replacement.* An individual observation in the original sample can appear more than once in the resample. <M, p.394>

For each of this resample the average is calculated. In the end we have thousands of averages.

The 95% confidence limits for the parameter "mean" are taken.
The lower bound : the average under which 2.5% of the resampled average lie
The upper bound: the average above which 2.5% of the simulated averages lie.

It is obvious that this new method has a lot of advantages especially for parameters others than the mean.
Also for non-normal data and for small samples (but ≥ 10) where we know nothing about their distribution.

For an easy and excellent introduction to bootstrap methods: Chapter 14 of Moore (2006) which is a separate pdf-file attached to the book.

The statistical reform in the 80s

Decision = Hypothesis tests contra Estimation = Confidence interval

In your scientific career you might increasingly be confronted with positions of either emphasizing or rejecting statistical hypothesis tests.

The conflict started around 1928 when Ronald Aylmer Fisher, the "father of modern statistics", and, among many others, the founder of the P value and the significance test concept, violently attacked the Type I and Type II error and hypothesis test theory of Jerzy Neyman and Egon Pearson (Moore, 2006, p. 439). Fisher's arguments and an excellent comparison of the two concepts you will find in Goodman (1993).

Null hypothesis tests (NHST) where established in medicine in the 50s as a new method to test the new drugs emerging in that period. "But just over a decade after the institutionalisation of NHST in medicine, its role in clinical trials was under scrutiny. Researchers began to worry that the technique was being misused and over-relied on; that statisticians, rather than physicians, had authority over the conclusions drawn from experiments (Cutler et al., 1966). Statistical reform had begun, and by the end of the 1980s strict editorial policies had profoundly changed the way results were reported, if not interpreted, in medicine. Effect size and CI reporting became routine reporting practice in many journals. " (Fiedler, 2006). A similar statistical reform was performed later in two other disciplines: ecology and psychology. Aiming a restricted use of hypothesis tests in favour of estimation and confidence intervals. (Fiedler, 2006)

Today, in medicine, more then 300 journals unite in CONSORT (2008) and formulate recommendations and checklists (see appendix), so that results should be accompanied by confidence intervals and actual P values rather than $P<0.05$.

In a "cost sensitive" discipline like epidemiology you will find quite radical rejection of hypothesis tests. Kenneth Rothman the editor of the journal Epidemiology informed: “when writing for Epidemiology, you can enhance your prospects if you omit tests of statistical significance… In Epidemiology, we do not publish them at all” (Rothman, 1998, p. 9). As a consequence the percentage of articles using mainly estimation and confidence intervals inccreased in such disciplines A survey reported a usage of 85% of all articles in 10 leading medical journals using confidence intervals in 2003. (Fiedler, 2006)

Nevertheless, modern textbooks in statistics generally cover decision theory so that you are able to understand respective articles and they also include all the hints and warnings you find in my text.

I will add a citation of Kenneth Rothman which summarizes quite drastically the arguments against hypothesis tests since Fisher's first warnings and makes it also clear why they fight so strictly:

> "In most scientific and public health settings, it is presumptuous if not absurd for an investigator to act as if the results of his or her study will form the sole basis for a decision. Such decisions are inevitably based on results from a collection of studies, and proper combination of the information from the studies requires more than just a classification of each study into "significant" or "not significant".
> Thus, degradation of information about an effect into simple dichotomy is counterproductive, even for decision making, and can be misleading."
> (Rothman, 2008, p. 155)

Modeling

Model

A construct or formulation that provides a description of the assumed structure of a set of data.

It involves a **set of assumptions** about distributions and relationships of random variables used to describe the data structure.

Distribution function F	Describes the regularities of a random variable: F(x) := Probability(X ≤ x)

Density function f	A positive function used to build F In case of discrete X it denotes Probability(X=x)

X discrete

Density function f for X	$0 \le f(j) \le 1$ $\sum_j f(j) = 1$ $f(i) := Pr(X = i)$
Distribution function F for X	$F(a) := Pr(X \le a) = \sum^{a} f(j)$

X continuous

Density function f for X	$f(x) \ge 0$ $\int_{-\infty}^{+\infty} f(x)dx = 1$ $Pr(a < X < b) = Pr(a \le X \le b =) = \int_a^b f(x)dx$
Distribution function F for X	$F(a) := Pr(X \le a) = \int_{-\infty}^{a} f(x)dx$

Density and Distribution Function

Examples: Discrete

Binomial distributions

Determined by two parameters: numbers of trials (n) and probability of a success per trial (p)

e.g. what is the probability of exact 2 times a "6" when casting a die 20 times:

$$f(2) = \binom{20}{2} \frac{1}{6}^2 \frac{5}{6}^{20-2} = 0.29$$

Mean = np

Standard deviation = np(1-p)

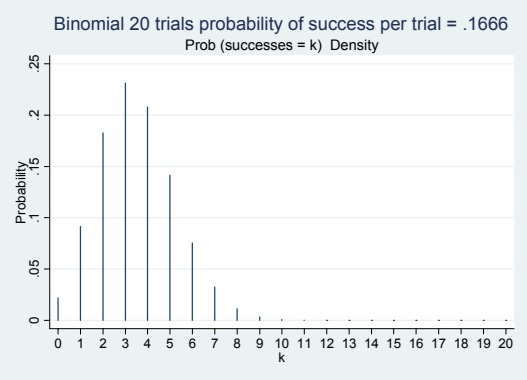
Density function

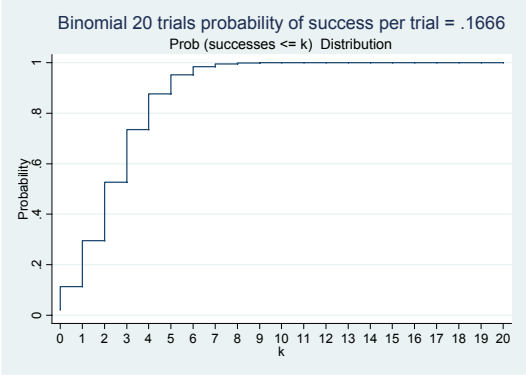
$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Distribution function

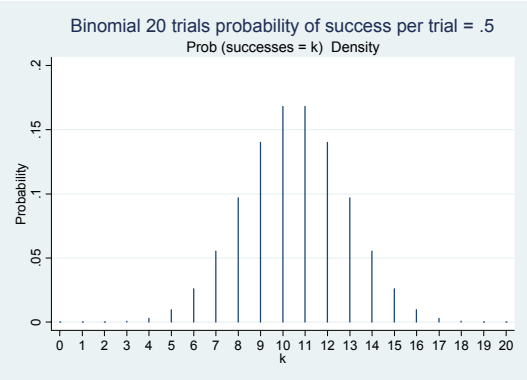
$$F(k) = \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j}$$

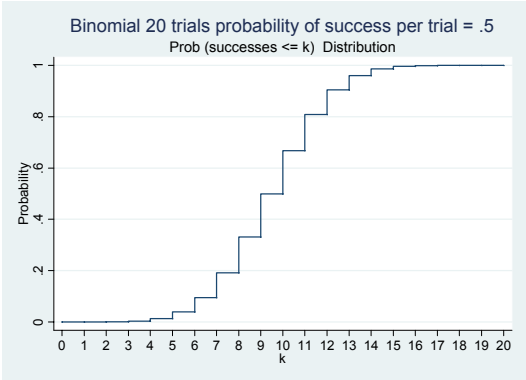
n=20 p=1/6



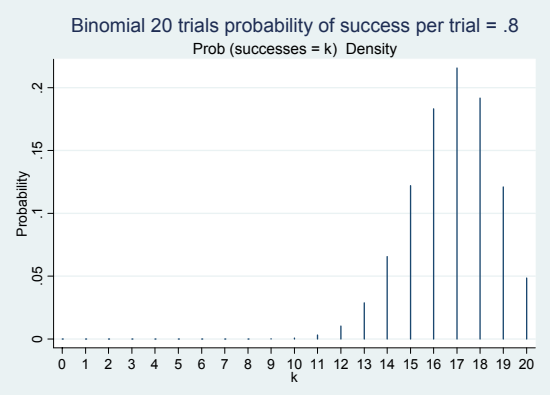


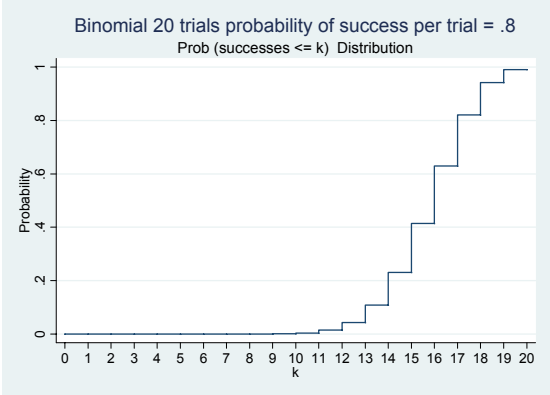
n=20 p=0.5





n=20 p=0.8





Density and Distribution Function

Examples: Continuous

Normal distributions N(μ;σ)

symmetric, unimodal, bell-shaped
Determined by two parameters Mean (μ) and Standard deviation (σ)

Mean = μ
Median = μ
Mode = μ
Standard deviation = σ
Skewness = 0
Kurtosis = 0

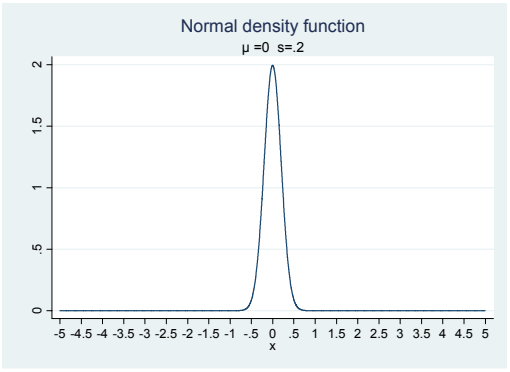
Density function

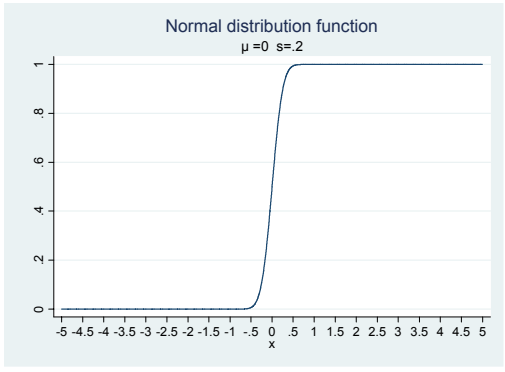
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Distribution function

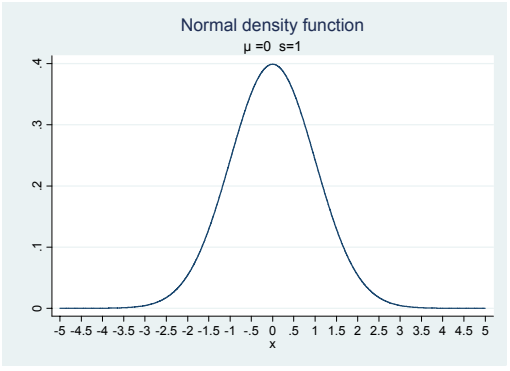
$$F(a) = \int_{-\infty}^{a} f(x)dx$$

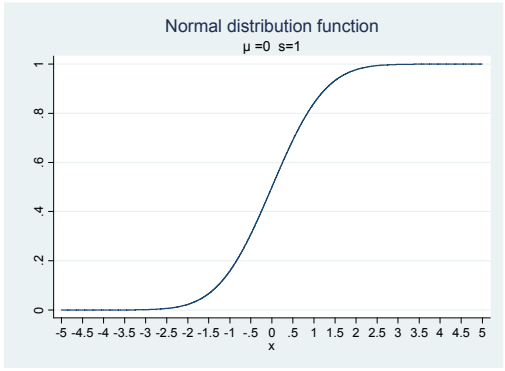
μ = 0 σ = 0.2



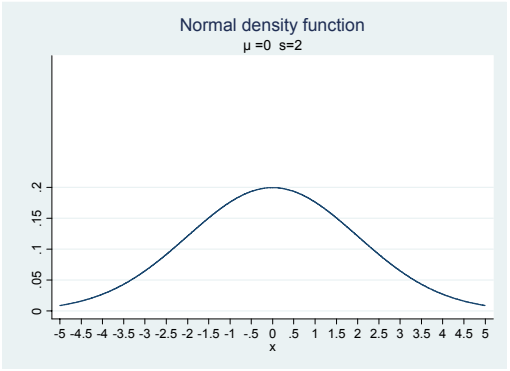


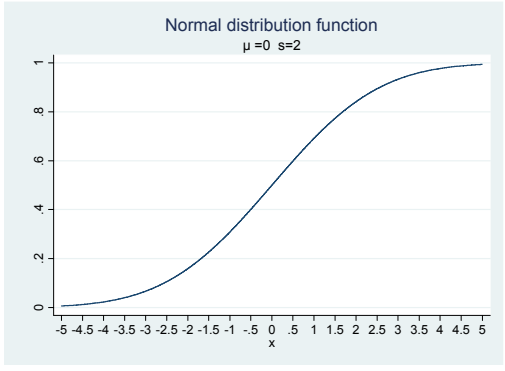
μ = 0 σ = 1





μ = 0 σ = 2





Density and Distribution Function

Examples: Continuous

Normal distributions N(μ;σ) cont.

1. **Central limit theorem**[12]: The distribution of the sample average of a large number of independent, identically-distributed variables (i.i.d.) random variables approaches the normal distribution with a mean μ and variance σ^2 / n irrespective of the shape of the original distribution

$$\text{Distribution}\left(\frac{1}{n}\sum_{i}^{n} X_i\right)_{n\to\infty} \to N$$

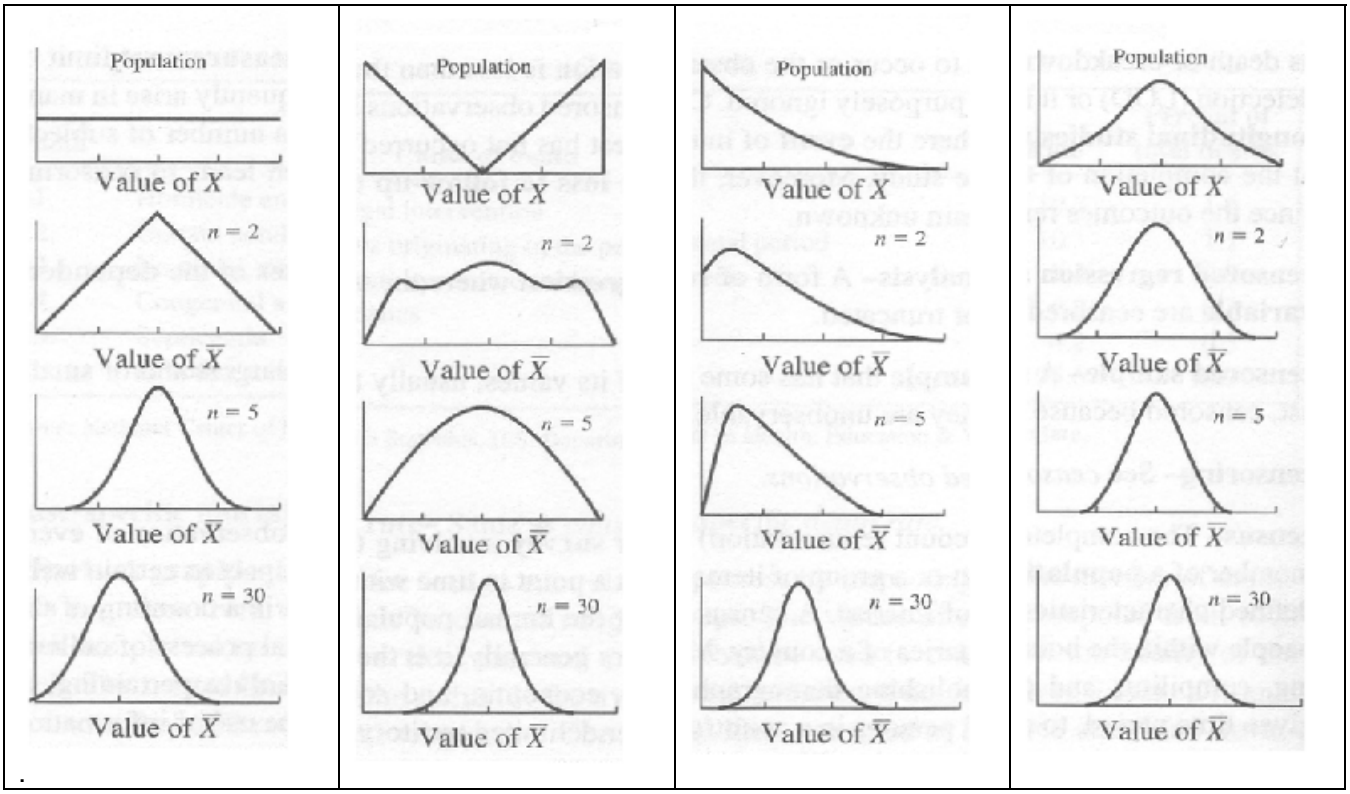
Which results in:

a. N(μ;σ) is a good description for **some** distributions of real data

   if many influences acting **additively and independently** to form the variable of interest

b. N(μ;σ) is the limit distribution for chance outcomes with many trials[13] (e.g. for binomial trials with large n)

c. Good approximation for other distribution in inference procedures if their distributions are roughly symmetric.[14]



.

Sahai (2002, p. 42)

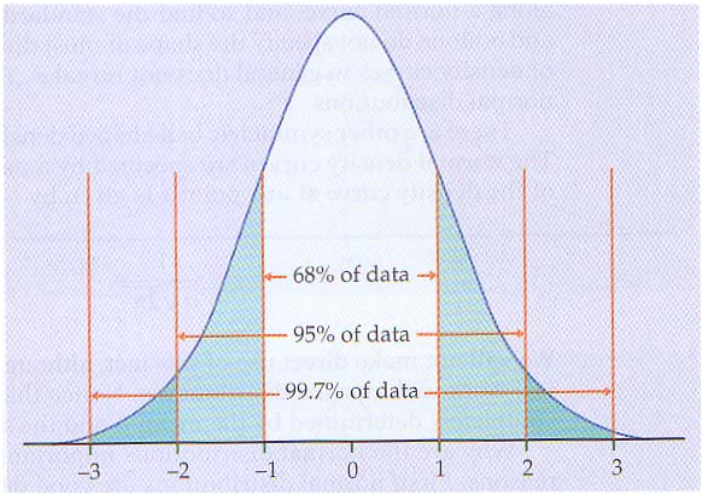
[12] Proved by Alexander Michailowitsch Ljapunow in 1901.
[13] First used by Abraham de Moivre in 1733 for binomial approximation if p=0.5 (coin tossing) then by Pierre-Simon Laplace in 1812 for general binomial distribution
[14] Moore (2006, p.71)

Density and Distribution Function

Examples: Continuous

Normal distributions N(μ;σ) cont.

2. The 68 95 99.7 Rule for normal distribution



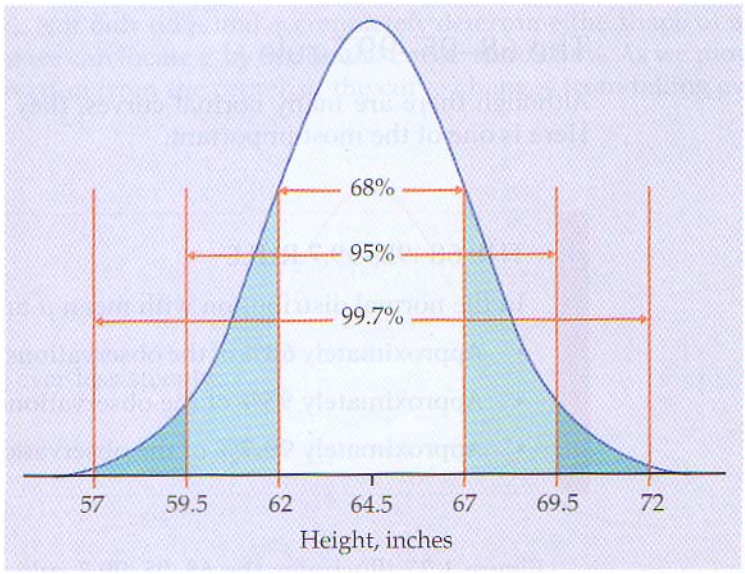
Moore (2006, p. 72)

3. The **Z-Transformation** converts any N(μ;σ) to the tabled standard Normal distribution N(0;1)[15]

$$Z = \frac{X - \mu}{\sigma}$$

X is N(μ;σ) ⟶ Z is N(0;1)

Example:
The distribution of heights of young women aged 18 to 24 is approximately normal with mean μ=64.5 inches and σ = 2.5 inches.



$$\text{Prob}(62 \le \text{Height} \le 67) = \text{Prob}\left(\frac{62 - 64.5}{2.5} \le \frac{\text{Height} - 64.5}{2.5} \le \frac{67 - 64.5}{2.5}\right) = \text{Prob}(-1 \le Z \le 1) = 0.68$$

[15] The distribution function of N(0;1) is called Φ <capital phi> and the density function of N(0;1) = φ <small phi>

Density and Distribution Function

Examples: Continuous

Lognormal distributions LN(μ;σ)

skewed, unimodal
Determined by two parameters μ and σ

Mean = $\exp(\mu+\sigma^2/2)$
Median = $\exp(\mu)$
Mode = $\exp(\mu-\sigma^2)$
Standard deviation = $\sqrt{[\exp(2\mu+\sigma^2)(\exp(\sigma^2)-1)]}$
Skewness = $\sqrt{[\exp(\sigma^2)-1)]}\,(\exp(\sigma^2)+2)$
Kurtosis = $\exp(4\sigma^2)+2\exp(3\sigma^2)+3\exp(2\sigma^2)-3$

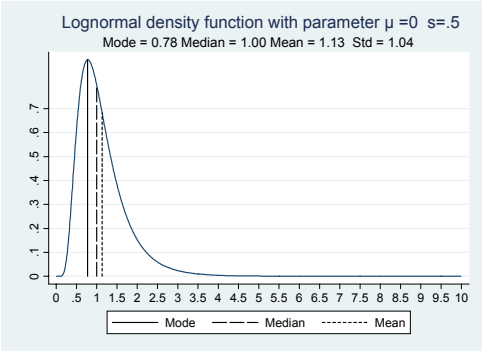
Density function

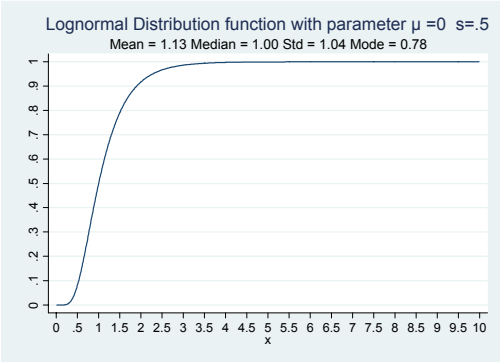
$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2}$$

Distribution function

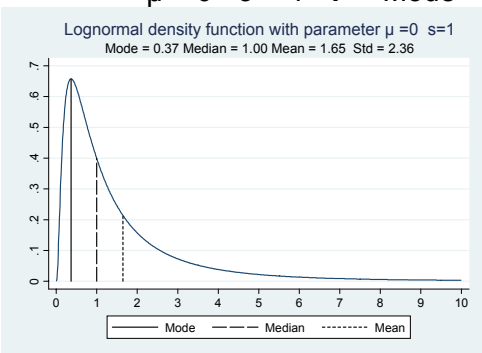
$$F(a)=\int_{-\infty}^{a}f(x)dx$$

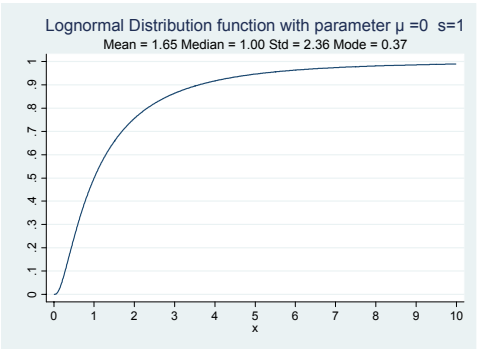
μ = 0 σ = 0.5 → Mode = 0.78 Median = 1.00 Mean=1.13 Std = 1.04



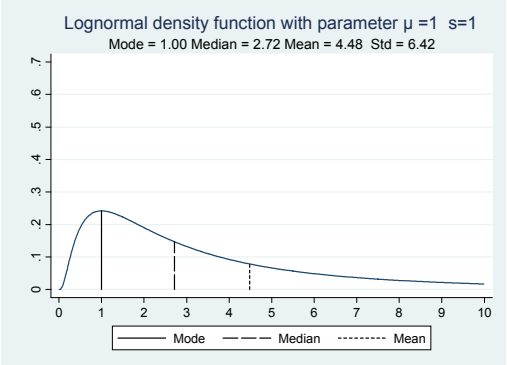


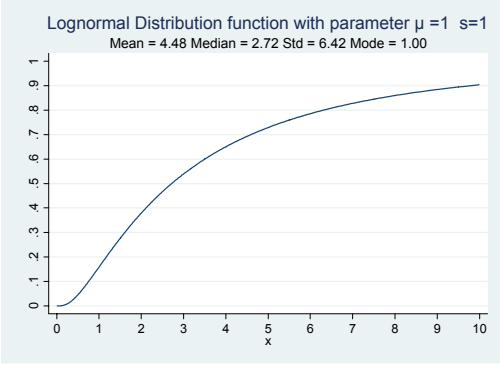
μ = 0 σ = 1 → Mode = 0.37 Median = 1.00 Mean=1.65 Std = 2.36





μ = 1 σ = 1 → Mode = 1.00 Median = 2.72 Mean=4.48 Std = 6.42





Density and Distribution Function

Examples: Continuous

Lognormal distributions LN(μ;σ) cont.

1. **Limit theorem:** "A log normal distribution results if the variable is the product of a large number of independent, identically-distributed variables in the same way that a normal distribution results if the variable is the sum of a large number of independent, identically-distributed" (Wolfram, 2001)

$$\text{Distribution}\left(\sqrt[n]{\prod_{i}^{n} X_i}\right) \underset{n\to\infty}{\to} \text{LN}$$

Therefore:

The class of lognormal distributions LN(μ;σ) serves as an appropriate model in
where many components influence or build up the variable of interest
multiplicatively (e.g. growth)
rather than additively

2. **X is Lognormal := ln(X) is Normal**

Therefore:

A lognormal distributed variable X can be transformed
via LN(X) into a normal distributed variable

Density and Distribution Function

Examples: Continuous

Lognormal distributions LN(μ;σ) cont.

Table 2. Comparing log-normal distributions across the sciences in terms of the original data. (Limpert et al., 2001, p. 347)

Discipline and type of measurement	Example	*n*	$\bar{x}$ *	s*	Reference
Geology and mining					
Concentration of elements	Ga in diabase	56	17 mg · kg^{-1}	1.17	Ahrens 1954
	Co in diabase	57	35 mg · kg^{-1}	1.48	Ahrens 1954
	Cu	688	0.37%	2.67	Razumovsky 1940
	Cr in diabase	53	93 mg · kg^{-1}	5.60	Ahrens 1954
	^{226}Ra	52	25.4 Bq · kg^{-1}	1.70	Malanca et al. 1996
	Au: small sections	100	(20 inch-dwt.)[a]	1.39	Krige 1966
	large sections	75,000	n.a.	2.42	Krige 1966
	U: small sections	100	(2.5 inch-lb.)[a]	1.35	Krige 1966
	large sections	75,000	n.a.	2.35	Krige 1966
Human medicine					
Latency periods of diseases	Chicken pox	127	14 days	1.14	Sartwell 1950
	Serum hepatitis	1005	100 days	1.24	Sartwell 1950
	Bacterial food poisoning	144	2.3 hours	1.48	Sartwell 1950
	Salmonellosis	227	2.4 days	1.47	Sartwell 1950
	Poliomyelitis, 8 studies	258	12.6 days	1.50	Sartwell 1952
	Amoebic dysentery	215	21.4 days	2.11	Sartwell 1950
Survival times after cancer diagnosis	Mouth and throat cancer	338	9.6 months	2.50	Boag 1949
	Leukemia myelocytic (female)	128	15.9 months	2.80	Feinleib and McMahon 1960
	Leukemia lymphocytic (female)	125	17.2 months	3.21	Feinleib and McMahon 1960
	Cervix uteri	939	14.5 months	3.02	Boag 1949
Age of onset of a disease	Alzheimer	90	60 years	1.16	Horner 1987
Environment					
Rainfall	Seeded	26	211,600 m^3	4.90	Biondini 1976
	Unseeded	25	78,470 m^3	4.29	Biondini 1976
HMF in honey	Content of hydroxymethylfurfurol	1573	5.56 g kg^{-1}	2.77	Renner 1970
Air pollution (PSI)	Los Angeles, CA	364	109.9 PSI	1.50	Ott 1978
	Houston, TX	363	49.1 PSI	1.85	Ott 1978
	Seattle, WA	357	39.6 PSI	1.58	Ott 1978
Aerobiology					
Airborne contamination by bacteria and fungi	Bacteria in Marseilles	n.a.	630 cfu m^{-3}	1.96	Di Giorgio et al. 1996
	Fungi in Marseilles	n.a.	65 cfu m^{-3}	2.30	Di Giorgio et al. 1996
	Bacteria on Porquerolles Island	n.a.	22 cfu m^{-3}	3.17	Di Giorgio et al. 1996
	Fungi on Porquerolles Island	n.a.	30 cfu m^{-3}	2.57	Di Giorgio et al. 1996
Phytomedicine					
Fungicide sensitivity, EC_{50}	Untreated area	100	0.0078 μg · ml^{-1} a.i.	1.85	Romero and Sutton 1997
Banana leaf spot	Treated area	100	0.063 μg · ml^{-1} a.i.	2.42	Romero and Sutton 1997
	After additional treatment	94	0.27 μg · ml^{-1} a.i.	>3.58	Romero and Sutton 1997
Powdery mildew on barley	Spain (untreated area)	20	0.0153 μg · ml^{-1} a.i.	1.29	Limpert and Koller 1990
	England (treated area)	21	6.85 μg · ml^{-1} a.i.	1.68	Limpert and Koller 1990
Plant physiology					
Permeability and solute mobility (rate of constant desorption)	*Citrus aurantium*/H_2O/Leaf	73	1.58 10^{-10} m s^{-1}	1.18	Baur 1997
	Capsicum annuum/H_2O/CM	149	26.9 10^{-10} m s^{-1}	1.30	Baur 1997
	Citrus aurantium/2,4–D/CM	750	7.41 10^{-7} 1 s^{-1}	1.40	Baur 1997
	Citrus aurantium/WL110547/CM	46	2.6310^{-7} 1 s^{-1}	1.64	Baur 1997
	Citrus aurantium/2,4–D/CM	16	n.a.	1.38	Baur 1997
	Citrus aurantium/2,4–D/CM + acc.1	16	n.a.	1.17	Baur 1997
	Citrus aurantium/2,4–D/CM	19	n.a.	1.56	Baur 1997
	Citrus aurantium/2,4–D/CM + acc.2	19	n.a.	1.03	Baur 1997
Ecology					
Species abundance	Diatoms (150 species)	n.a.	12.1 i/sp	5.68	May 1981
	Plants (coverage per species)	n.a.	~0.4%	7.39	Magurran 1988
	Fish (87 species)	n.a.	2.93%	11.82	Magurran 1988
	Birds (142 species)	n.a.	n.a.	33.15	Preston 1962
	Moths in England (223 species)	15,609	17.5 i/sp	8.66	Preston 1948
	Moths in Maine (330 species)	56,131	19.5 i/sp	10.67	Preston 1948
	Moths in Saskatchewan (277 species)	87,110	n.a.	25.14	Preston 1948
Food technology					
Size of unit (mean diameter)	Crystals in ice cream	n.a.	15 μm	1.5	Limpert et al. 2000b
	Oil drops in mayonnaise	n.a.	20 μm	2	Limpert et al. 2000b
	Pores in cocoa press cake	n.a.	10 μm	1.5–2	Limpert et al. 2000b

Table 2. (continued from previous page)

Discipline and type of measurement	Example	*n*	$\bar{x}$ *	s*	Reference
Linguistics					
Length of spoken words in phone conversation	Different words	738	5.05 letters	1.47	Herdan 1958
	Total occurrence of words	76,054	3.12 letters	1.65	Herdan 1958
Length of sentences	G. K. Chesterton	600	23.5 words	1.58	Williams 1940
	G. B. Shaw	600	24.5 words	1.95	Williams 1940
Social sciences and economics					
Age of marriage	Women in Denmark, 1970s	30,200	(12.4)[a] 10.7 years	1.69	Preston 1981
Farm size in England and Wales	1939	n.a.	27.2 hectares	2.55	Allanson 1992
	1989	n.a.	37.7 hectares	2.90	Allanson 1992
Income	Households of employees in Switzerland, 1990	1.7x10^6	sFr. 6,726	1.54	Statistisches Jahrbuch der Schweiz 1997

[a] The shift parameter of a three-parameter log-normal distribution.

Notes: n.a. = not available; PSI = Pollutant Standard Index; acc. = accelerator; i/sp = individuals per species; a.i. = active ingredient; and cfu = colony forming units.

Model

Relation
Linear Model

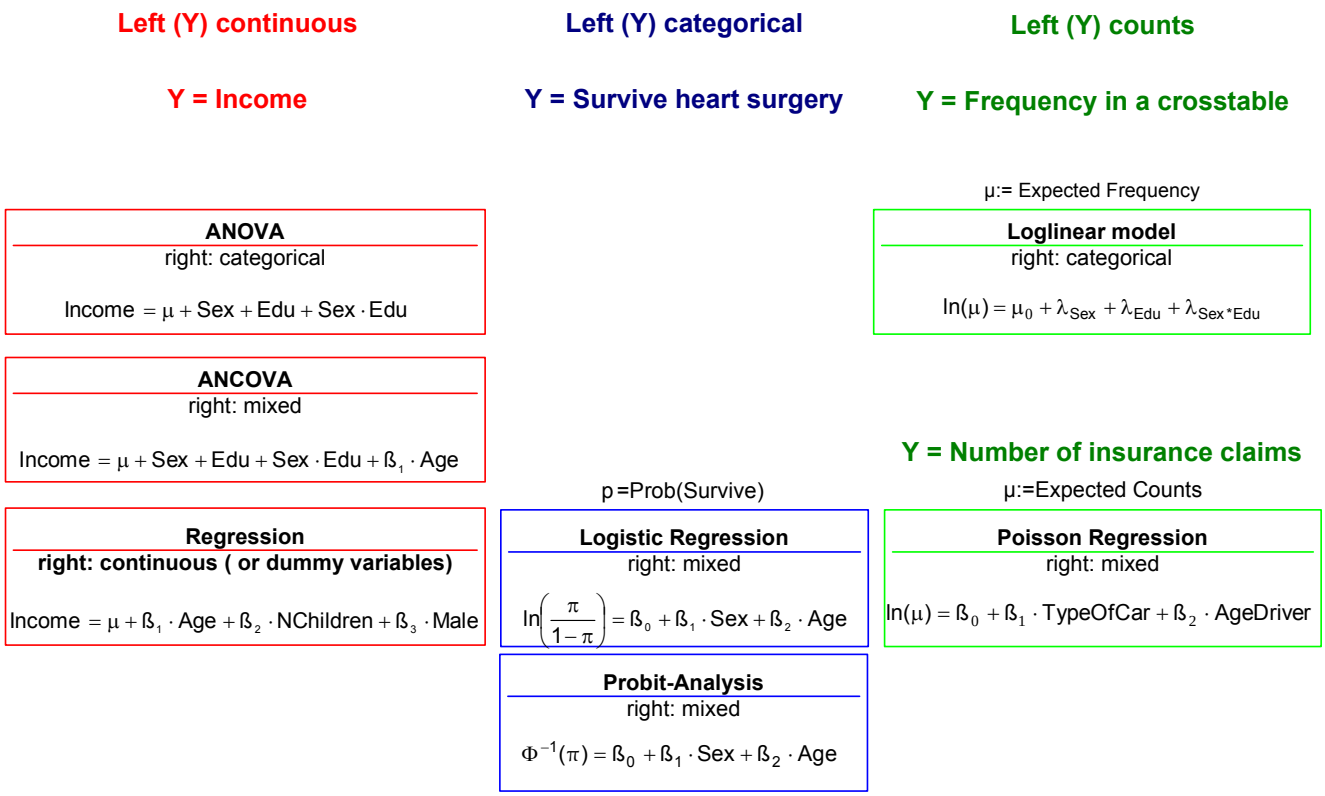
"There is no way to infer the 'right' model from the data unless there is strong prior theory to limit the univers of possible models." (Freedman, 2006, p.193)

Linear model

The most frequently used models in statistics are linear models. They are used to test and estimate effects of factors.

Here is an overview of the most used linear models for various situations

Figure Most frequently used linear models with model equations in infomal style



In 1972, Nelder and Wedderburn found that all these, and even more models can be fitted and tested via one technique: the maximum likelihood method (ML). They called this class of models the **Generalized Linear Model** (GZLM). But, because ML estimates are only asymptotically unbiased and normally distributed (e.g. see Freedman, 2006 p.113), the models on the left side (Y continuous) are mostly solved with the older OLS-techniques which results in best linear unbiased estimators (BLUE) even for smallest sample sizes. These models on the left side are summarized under the **General Linear Model** (GLM).

The abbrevations are not consitent as, for instance, Nelder abbreviate his generalized linear model with GLM (McCullagh & Nelder, 1989, p.21).

Model
Relation **Linear Model**

Generalized linear model (P values and confidence are asymptotically correct) if

Standard form:

$g(\mu) = g(\mu(x)) = g(E(Y|x)) = ß_0 + ß_1x_1 + ß_2x_2 + \ldots + ß_kx_k$

$g(\mu) = X ß$ μ (nx1), X (nxk) ß (kx1)

ANOVA, ANCOVA and Regression

Observations independently Normal distributed

Logistic Regression, Probit:

Observations independently Binomial distributed

Poisson Regression

Observations independently Poisson distributed

General linear model (P values and confidence are correct) if

Standard form:

$y_i = ß_0 + ß_1x_{1i} + ß_2x_{2i} + \ldots + ß_kx_{ki} + e_i$ i=1,…,n e_i are i.i.d.

$y = X ß + e$ y (nx1), X (nxk), ß (kx1), e (nx1)

ANOVA

$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + e_{k(ij)}$ see Underwood (1998, p. 306)

where Y_{ijk} is the kth replication of factorlevel i of A and factorlevel j of B

Here A_i is called the effect of A_i (i.e. $\mu_i - \mu$) with μ_i = Mean of factor level i of A so that the the model can also be written as

$Y_{ijk} = \mu_{ij} + e_{ijk}$ see Moore (2006, p. 776)

ANCOVA

$Income_i = \mu + \mu_{Sex} + \mu_{Edu} + \mu_{Sex}{}^*\mu_{Edu} + ß Age_i + e_i$ i=1,…,n

Regression

$Income_i = ß_0 + ß_1 Age_i + ß_2 Children_i + e_i$ i=1,…,n

Appendix

Chapter "Statistical methods" from the CONSORT statement

http://www.consort-statement.org/index.aspx?o=1029

Statistical methods

12(a) Statistical methods used to compare groups for primary outcome(s)

Example

"All data analysis was carried out according to a pre-established analysis plan. Proportions were compared by using Chi-squared tests with continuity correction or Fisher's exact test when appropriate. Multivariate analyses were conducted with logistic regression. The durations of episodes and signs of disease were compared by using proportional hazards regression. Mean serum retinol concentrations were compared by t test and analysis of covariance ... Two sided significance tests were used throughout." (125)

Explanation

Data can be analyzed in many ways, some of which may not be strictly appropriate in a particular situation. It is essential to specify which statistical procedure was used for each analysis, and further clarification may be necessary in the results section of the report.

Almost all methods of analysis yield an estimate of the treatment effect, which is a contrast between the outcomes in the comparison groups. In addition, authors should present a confidence interval for the estimated effect, which indicates a range of uncertainty for the true treatment effect. The confidence interval may also be interpreted as the range of values for the treatment effect that is compatible with the observed data. It is customary to present a 95% confidence interval, which gives the range of uncertainty expected to include the true value in 95 of 100 similar studies.

Study findings can also be assessed in terms of their statistical significance. The P value represents the probability that the observed data (or a more extreme result) could have arisen by chance when the interventions did not differ. Actual P values (for example, P = 0.003) are preferred to imprecise threshold reports (P < 0.05) (46, 127).

Standard methods of analysis assume that the data are "independent". For controlled trials, this usually means that there is one observation per participant. Treating multiple observations from one participant as independent data is a serious error; such data arise when outcomes can be measured on different parts of the body, as in dentistry or rheumatology. Data analysis should be based on counting each participant once (128, 129) or should be done by using more complex statistical procedures (130). Incorrect analysis of multiple observations was seen in 123 (63%) of 196 trials in rheumatoid arthritis (28).

12(b) Methods for additional analyses, such as subgroup analyses and adjusted analyses

Examples

"Proportions of patients responding were compared between treatment groups with the Mantel-Haenszel Chi-squared test, adjusted for the stratification variable, methotrexate use" (80).

"…it was planned to assess the relative benefit of CHART in an exploratory manner in subgroups: age, sex, performance status, stage, site, and histology. To test for differences in the effect of CHART, a chi-squared test for interaction was performed, or when appropriate a chi-squared test for trend" (131).

Explanation

As is the case for primary analyses, the method of subgroup analysis should be clearly specified. The strongest analyses are those based on looking for evidence of a difference in treatment effect in complementary subgroups (e.g., older and younger participants), a comparison known as a test of interaction (132, 133). A common but inferior approach is to compare P values for separate analyses of the treatment effect in each group. It is incorrect to infer a subgroup effect (interaction) from one significant and one non-significant P value (134). Such inferences have a high false-positive rate.

Because of the high risk for spurious findings, subgroup analyses are often discouraged (14, 135).

Post hoc subgroup comparisons (analyses done after looking at the data) are especially likely not to be confirmed by further studies. Such analyses do not have great credibility.

In some studies, imbalances in participant characteristics (prognostic variables) are adjusted for by using some form of multiple regression analysis.

Although the need for adjustment is much less in RCTs than in epidemiological studies, an adjusted analysis may be sensible, especially when one or more prognostic variables seem important (136).

Ideally, adjusted analyses should be specified in the study protocol. For example, adjustment is often recommended for any stratification variables (see item 8(b)). In RCTs, the decision to adjust should not be determined by whether baseline differences are statistically significant (133, 137) (see item 16). The rationale for any adjusted analyses and the statistical methods used should be specified.

Authors should clarify the choice of variables that were adjusted for, indicate how continuous variables were handled, and specify whether the analysis was planned or suggested by the data (Müllner M, Matthews H, Altman DG. Reporting on statistical methods to adjust for confounding: a cross sectional survey. Submitted for publication.). Reviews of published studies show that reporting of adjusted analyses is inadequate with regard to all of these aspects (138-140).

Page last edited: 27 July 2007

Citations:

Freedman (2006):

"The goal of empirical research is - or should be - to increase our understanding of the phenomena, rather than displaying our mastery of technique." (p. 200)

"In the social and behavirol sciences, far-reaching claims are often made for the superiority of advance quantitative methods - by those who manage to ignore the far-reaching assumptions behind the model" (p. 200)

"Many statisticians find it surprising that regression and allied techniques (like logistic regression) are commonly used in the social and life sciences to infer causation from observational data, with qualitative inference perhaps more common than quatitative: X causes (or doesn't cause) Y, the magnitude of the effect being of lesser interest. Eyebrows are sometimes raised about the whole idea of causation…" (p.147)

"There is no way to infer the "right" model from the data unless there is a strong prior theory to limit the universe of possible models" (p. 193)

"There is no one, therefore, so far human qualities go whom it would be safer to trust with black magic. That there is anyone I would trust with it at the present stage, or that this brand of statistical alchemy is ripe to become a brachn of science" Keynes about Tinbergen, one of the pioneers of econometric modeling. (Keynes 1940, p. 156 cited in Freedman, 2006, p. 195)

Other

"There is no correct method for inference from data haphazardly collected with bias of unknown size. Fancy formulas cannot rescue badly produced data." (Moore, 2006, p. 393)

"Attempts to separate the effects of two confounded variables by observation alone are always silly (whoever does them!)" (Underwood, 1989, p. 439)

References

Suggested textbooks

Moore DS, McCabe GP (2006). *Introduction to the practice of statistics 5. ed.*. Freeman, New York.

Utts JM (2004). *Seeing Through Statistics*. Thomson, Belmont.

van Belle G (2008). *Statistical Rules of Thumb*. Wiley, New York.

Good PI, Hardin JW (2006). *Common Errors in Statistics (and How to Avoid Them)*, Wiley, New York.

Altman D, Gardner M (2000). *Statistics with confidence*. British Medical Journal Books, London.

Rothman KJ (2008). *Modern epidemiology*. 3rd edition, Lippincott Williams & Wilkins, Philadelphia.

Freedman D (2006). *Statistical Models: Theory and Practice.* Cambridge University Press.

Levy PS, Lemeshow S (2008). *Sampling of Populations*. New York: Wiley.

Used in this overview

Armitage P, Colton T (2005). *Encyclopedia of Biostatistics*.Vol.1-8, Wiley, New York.

McCullagh P (2002). What is a Statistical Model? The Annals of Statistics, Vol. 30, No. 5, pp. 1225-1267.

Besag J et al. (2002). What is a Statistical Model? Discussion. The Annals of Statistics, Vol. 30, No. 5, pp. 1267-1310.

Limpert E et al. (2001). Log-normal Distributions across the Sciences: Keys and Clues, Bioscience, 51 (5), p. 341–352 < http://stat.ethz.ch/~stahel/lognormal/bioscience.pdf>.

Goodman SN (1993). p Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate. Am J Epidemiol 1993;137:485-96.

Fidler F (2006). Should Psychology abandon p values and teach confidence intervals instead? Evidence-based reforms in statistics education. Working co-operatively in statistics education. Proceedings of ICOTS-7, Seventh International Conference on Teaching Statistics. Salvador, Brazil. < http://www.stat.auckland.ac.nz/~iase/publications/17/5E4_FIDL.pdf >

Rothman KJ (1988). *Modern epidemiology*. 2nd edition, Lippincott Williams & Wilkins, Philadelphia.

The CONSORT Statemant (2008) < http://www.consort-statement.org/ >

Underwood A (1998). Experiments in ecology. Cambridge University Press.

Sahai H, Khurshid A (2002). *Dictionary of Statistics*. McGraw-Hill, Boston.

Other Internet Resources

Wikipedia Statistics Portal < http://en.wikipedia.org/wiki/Statistics >.

Wolfram Math World < http://mathworld.wolfram.com/ >.

PEAS (2006) Practical Exemplars on the Analysis of Surveys. Napier University, Edinburgh. < http://www2.napier.ac.uk/depts/fhls/peas >.

Gardner M, Altman, D (1989) CIA Confidence Interval Analysis Version 1, < http://www.guidoluechters.de/References/Programs/cia.exe >. Version 2.1.2 in Altman D, Gardner M (2000).

Historical:

Tukey JW (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading.

Bowley AL (1901). *Elements of Statistics*, (4th edition in 1920)[16]. London.

[16] "The *Elements of Statistics* is generally regarded as the first English-language statistics text-book" Wikipedia